Machine Learning

Objetivo

El objetivo de la práctica es simple: abordar un problema de Machine Learning realista siguiendo la metodología y buenas prácticas explicadas durante las clases teóricas. Por tanto, en estas instrucciones no se especifican los pasos exactos que el alumno tiene que llevar a cabo para realizar esta tarea con éxito; es parte del trabajo aplicar las técnicas de procesamiento/transformación de variables que mejor se adecúen al problema, identificar los modelos que proporcionen prestaciones óptimas, las variables potencialmente más relevantes y la métrica adecuada para contrastar los distintos modelos. Aún así, se proporciona una pequeña guía de los pasos necesarios. Las posibilidades son amplias, así que es recomendable abordar una aproximación incremental: comenzar por soluciones sencillas para progresivamente aumentar la complejidad de las técnicas utilizadas.

A diferencia de los datasets utilizados en las clases, este está compuesto por datos reales, es decir, precisa de un análisis y limpieza mayores. Por el mismo motivo no se pretende obtener unos resultados espectaculares, es suficiente con que sean decentes; se valorará mucho más que el proceso seguido tenga sentido y no contenga errores graves de concepto.

Conjunto de datos

El conjunto de datos escogido para la práctica es un dataset que contiene un extracto de incidencias relacionadas con ciberataques a distintos sistemas. Contiene distintas columnas de las cuales tres son target o columnas a predecir. Estas columnas son “Action Taken”, “Severity Level” y “Attack Type”.

Solamente es obligatorio predecir para “Action Taken” pero la predicción de las otras dos se valorará positivamente.

No todas las columnas serán útiles. Algunas se podrán usar sin apenas tratamiento, otras necesitan de alguna transformación básica como imputación de valores nulos y otras necesitarán de un tratamiento más profundo.

No es necesario usar todas, es más importante usar las columnas analizadas en detalle y que se esté seguro de que son útiles a usar todas ciegamente.

Tarea

Es un problema de clasificación:

1. Preparación de datos: División train/test
2. Análisis exploratorio, por ejemplo:
    a. Head, describe, dtypes, etc.
    b. Outliers
    c. Correlación
3. Preprocesamiento:
    a. Eliminación de variables, mediante selección (random forest/Lasso), alta correlación, alto porcentaje de missings, o el método que se considere oportuno.
    b. Generación de variables
4. Modelado:
    a. Cross validation
    b. Evaluación; mejor si lo hacéis de más de un modelo, porque así podéis comparar entre ellos.
5. Conclusión: escrita, no numérica; un par de líneas es más que suficiente.

Es más importante una explicación correcta sobre qué puede estar pasando con los datos y los modelos entrenados que una práctica que contenga todas las técnicas dadas durante el módulo.

Se valorará más positivamente un entrenamiento correcto y robusto que un modelo que tenga mejores métricas que la media. Se valorará negativamente fallos graves de concepto que pongan de manifiesto que no se han comprendido las bases de cómo se trabaja cuando uno se enfrenta a un problema con los datos.

Si habéis hecho todas las tareas necesarias, no os preocupéis si durante la práctica llegáis a un impasse y no sabéis cómo continuar. Dedicadle unos momentos a pensar que puede estar pasando y escribidlo en la conclusión.

Recordad que lo importante es demostrar que sabéis entender y trabajar en un problema de ML.

Modo de entrega

Hay que realizar la práctica en Python y subirla al formulario de entrega que se facilitará más
adelante. No basta con escribir código; hay que explicar lo que se ha hecho de forma suficientemente detallada, preferiblemente con gráficas y/o comentarios en markdown (o en el propio código Python, no hay problema). La estructura del proyecto es indiferente, puede ser en un archivo .py o en cuadernos de Jupyter .ipynb.